Pre-Grant and Grant Yellow Book 2

United States Patents
(Grants and Applications)
delivered as
CCITT Group 4 Facsimile Images
with TIFF Headers
on DLT Cartridges

MARCH 1, 2004

United States Patent & Trademark Office
Office Electronic Information Products

1 Background

The USPTO implementation of the World Intellectual Property Office (WIPO) Standard ST.33 is known as Yellow Book. ST.33 provides a proprietary header for CCITT Group 4 compressed raster images.

Yellow Book 2 uses a TIFF header for CCITT Group 4 compressed raster images of the pages in the document, accompanied by an XML instance with additional metadata for each document. The documents published in any given week are archived using standard UNIX "tar" (Tape ARchive) software and written to a DLT tape cartridge. Yellow Book 2 is based on WIPO Standards ST.33, ST.35, and current USPTO practice.

2 Summary

Yellow Book 2 consists of United States patent grant publications and patent application publications, as well as certificates of correction and reexamination certificates therefore, delivered as CCITT Group 4 facsimile images enclosed in TIFF headers. Each page of a document is in a single TIFF file. The files are organized into directories, one directory per document. Each DLT cartridge contains an additional Tape Content List (TCL) file that identifies all document numbers that reside on the DLT cartridge.

3 Organization of DLT Contents

3.1 Tape ID File

The Tape ID file is the first archived file appearing on the tape. The Tape ID file contains no data. The name of the Tape ID file is the volume serial number of the tape cartridge:

`Xnnnnn.tid`

where `X` will be an upper case alpha character that identifies the tape series and `nnnnn` will the tape numeric number right justified with leading zeros.

Tape Series Codes	
A	Patent Grants prior to June 4, 2002 – 3480 Magnetic Tape Cartridge
B	Rescans

D	Patent Grants for Dissemination
G	Patent Grants from data capture contractor
P	Patent Applications from data capture contractor
R5	Certificates of Extension
T	Patent Applications for Dissemination
Z	Certificates of Corrections from data capture contractor

3.2 Tape Content List (TCL) File

The file name for TCL file will be:

`Yyyymmdd.contents`

For weekly publication of grants and applications, `yyyy` is the year, `mm` is the month and `dd` the day of the month, representing the issue/publication date of the images on the tape. For all other types of content, `yyyymmdd.contents` is the date the tape was written.

The TCL will list each document stored on the tape. Each entry in the TCL will contain the document ID, kind code, issue/publication date, and page count. The data fields will be tab delimited. Each entry in the TCL will be delimited with a new line character. The data will be in ASCII format.

Patent Grants and Certificates of Correction :

```
C6303038  C1  20040106  2
D0485873  S1  20040106  6
H0002097  H1  20040106  4
PP014508  P2  20040106  3
RE038412  E1  20040106  19
06300737  B1  20040106  11
06300737  X6  20040106  1
06303038  B1  20040106  7
```

Patent Applications:

```
US20010025645A1  A1  20040101  10
US20010025646A1  A1  20040101  6
US20010025647A1  A1  20040101  12
US20010025648A1  A1  20040101  15
```

This file will be the second archived file on the tape.

3.3 Images and Metadata Archive File

Documents are grouped under a directory that is at the root of the tape (tar) directory structure. The directory will be named one of the following:

`yyyy-ww`

where `yyyy` is the year and `ww` is two digit week of the year that the documents where created or modified.

`Uyyyymmdd-yyyymmdd`

where `U` identifies a tape which contains documents that were created or modified within a specified date range. The dates are in the following format were `yyyy` is a year, `mm` is a month, and `dd` is a day.

`Pyyyymmdd-yyyymmdd`

where `P` identifies a tape which contains documents that were issued/published within a specified date range. The dates are in the following format were `yyyy` is a year, `mm` is a month, and `dd` is a day.

`DocID-DocID`

where a tape contains documents that have document IDs between the two document IDs appearing in the root directory name.

The directory structure, images, and metadata make up the third archive file on the tape.

4 Document Page Images

4.1 Directory Structure

A directory structure will be created, one subdirectory per document, to store the page images (TIFF files) and the document-level metadata (XML instance file). The hierarchy of the directory structure containing patent documents will be:

Root_Directory_Name/12/345/678

12345678 represents the 8-position patent document number. This structure is intended to ensure that there are no more than 1,000 subdirectories in a directory.

The hierarchy of the directory structure containing publication documents will be:

Root_Directory_Name/US/YYYY/1234/567/KC

US is the country code. YYYY represents the publication year of the application. 123/4567 represents the 7-position serial number of the application. KC represents the kind code of the application.

The number of files in a subdirectory reflects the number of pages in a document plus the metadata.

4.2 Page Image Files

The TIFF file name for each page image will be:

????????.tif

where ???????? is an eight-character field containing the page number, right-aligned and left-padded with zeros. The page number represents the sequence of the image page within the document.

4.3 TIFF Header Contents

The TIFF header of each page image contains standard TIFF header tags and the following tags derived from WIPO Standard ST.35. Tag 50560 has been added and tags 269, 274, 306, and 999 have been modified from the original in ST.35.

ID	Meaning of ite	Data type	Length	Value or pointer	Remarks
254	New subfile type	4	1	0	Indicates that it is a full resolution image. Default value 0.
255	Old subfile type	3	1	1	For compatibility reasons still available.
256	Width of image	3	1	number	In pixels (X direction).
257	Length of image	3	1	number	In pixels (Y direction).
258	Bits per sample	3	1	1	Black and white, 1 bit per sample.
259	Compression method	3	1	4	ITU-T (CCITT) Fax Group 4.

ID	Meaning of ite	Data type	Length	Value or pointer	Remarks
262	Photometric interpretation	3	1	0	Minimum value (0) is white, maximum value (1) is black.
266	Fill order	3	1	1	Left to right.
269	**Document name**	**2**	**24**	**xx**	**xx is a pointer to the full document number (based on WIPO Standard ST.14) as follows: Publishing office country code (2 positions); Document number (12 positions, right justified, left padded with zeros); Kind code (two positions); Date (eight positions, CCYYMMDD).**
270	Image description	2	9	xx	xx is a pointer to the image identification, which consists of a page number (4 positions) and a frame number (4 positions) + 1 end byte.
273	Strip offset	4	1	xx	xx is a pointer to the start of the image data belonging to this directory.
274	**Orientation**	**3**	**1**	**0**	**Rotation or orientation of image: 0 = portrait (default); 1 = landscape**
277	Samples Per Pixel	3	1	1	Black and white.
278	Rows per strip	4	1	number	Number of rows (equal to tag 257, height in pixels).
279	Strip byte count	4	1	number	Number of bytes of image data in uncompressed form.
280	Min sample value	3	1	0	
281	Max sample value	3	1	1	
282	X resolution	5	1	xx	xx is a pointer to the field containing the numerator of the resolution in pixels in x direction, which is 4 bytes long. The value of this field is 300. The denominator follows this field immediately and is also 4 bytes long. The value of this field is 1. The result is a value of 300 DPI in x direction.
283	Y resolution	5	1	xx	Resolution in y direction, see tag 282 for explanation. The value is 300 DPI.
293	Group 4 options	4	1	0	Compressed in ITU-T (CCITT) Gr 4 format.
296	Resolution unit	3	1	2	Inches.
306	**Date time**	**2**	**20**	**xx**	**xx is a pointer to the field containing the Date (YYYY:MM:DD) and the Time (HH:MM:SS). This is the creation date of the TIFF header.**
999	Miscellaneous	2	253	xx	Private field. By default, this field is blank.
50560	**Original content type**	**3**	**1**	**0**	**0 = text or black & white drawing (default); 1 = grayscale drawing or photograph; 2 = color drawing or photograph**

4.4 Metadata File

For each document there will be a metadata file that is an instance of the following document type definition. The file name of the metadata file for each document will be us-patent-image.xml.

```
<!--Document Type Definition for metadata to accompany facsimile images of
United States patents.
Reference this DTD as PUBLIC "-//USPTO//DTD us-patent-image v1.0 2002-06-
04//EN"
Alias: Yellow Book 2 (YB2)
Contact: Ed Johnson
Information Products Division
U.S. Patent and Trademark Office
Crystal Park 3, Suite 441
Washington, DC 20231
vox: 703-306-2621
fax: 703-306-2737
ed.johnson@uspto.gov

******** Revision History ********
2003-06-10 Barry Frank
. Changed all references of element name "drawup" to "scan-date".
. Changed all references of element name "withdrawn-flag" to "withdrawn-
indicator".
. Changed all references of element name "start" to "begin". Also changed
comments referring to start
..        to refer to begin.
2003-03-28 Barry Frank
. Added bib-pages?,abstract-pages?,drawings-pages?,description-pages?,claims-
pages? to
..      the reexamination-certificate element.
. Removed the ? from the related-document element in the certificate-of-
correction
..      and reexamination-certificate elements. (A related document must be
present)
2002-06-18 Bruce B. Cox
. Final version 1.  Added withdrawn as valid status type.
2002-06-04 Bruce B. Cox
. Final draft of version 1. Eliminated page metadata content model and revised
document metadata
.. content model. All page-specific information now in TIFF header, for a
description of which, see YB2
.. specification.
2002-05-10 First public draft.
****** End Revision History ******

-->
<!ELEMENT us-patent-image  (patent-metadata?,certificate-of-correction*,
               reexamination-certificate*) >

<!ATTLIST us-patent-image
               file  CDATA    #REQUIRED
               file-type (tiff)   #FIXED "tiff"
               date-produced  CDATA    #REQUIRED
               lang  CDATA    #REQUIRED
               dtd-version  CDATA    #IMPLIED
```

```
                status  CDATA    #IMPLIED
                country  CDATA    #FIXED "us" >

<!--For both US Patent Application Publications and US Patent Grants.
Data-capture contractor will use patent-metadata for all deliverables (grants,
applications, certificates of correction, and reexamination certificates).
Dissemination products, however, will use patent-metadata, certificate-of-
correction, and reexamination-certificate appropriately.-->
<!ELEMENT patent-metadata  (full-document-number,document-id,page-count,scan-
date,
                record-status,related-document?,withdrawn-indicator?,missing-
pages?,
                bib-pages?,abstract-pages?,drawings-pages?,description-pages?,
                claims-pages?,certificate-of-correction-pages?,reexamination-
pages?) >

<!--Begin and End indicate the first and last pages of just this one
certificate of correction relative to the entire document-->
<!ELEMENT certificate-of-correction  (document-id,page-count,scan-date,record-
status,
                related-document,missing-pages?,begin,end) >

<!--Begin and End indicate the first and last pages of just this one
reexamination certificate relative to the entire document-->
<!ELEMENT reexamination-certificate  (document-id,page-count,scan-date,record-
status,
                related-document,missing-pages?,begin,end,bib-pages?,abstract-
pages?,drawings-pages?,description-pages?,
                claims-pages?) >

<!--The complete document identification, arranged for display, as in ST.14-->
<!ELEMENT full-document-number  (#PCDATA) >

<!--Document identification refers to patents and patent applications only.
See WIPO ST.14-->
<!ELEMENT document-id  (country,doc-number,kind,name?,date?) >

<!ATTLIST document-id
                lang  CDATA    #IMPLIED  >

<!--Total number of image pages in the document.-->
<!ELEMENT page-count  (#PCDATA) >

<!--Date that page image(s) were created.-->
<!ELEMENT scan-date  (date) >

<!--New = page images of a new publication
Rescan = some or all of the image pages have been replaced with corrected
images, or addition of missing pages
Delete = all images of the referenced document should be deleted-->
<!ELEMENT record-status   EMPTY  >

<!ATTLIST record-status
             value (new | rescan | retro | delete | withdrawn)   #REQUIRED  >
```

```
<!--If the document is a reissue patent, this is the number of the original
document. If the document is a certificate of correction, this is the number
of the corrected document.-->
<!ELEMENT related-document  (doc-number) >

<!--Indicates that the document has been withdrawn.-->
<!ELEMENT withdrawn-indicator   EMPTY  >

<!--Contains a list of missing pages, comma separated. If the element is
present but no page numbers are present, there are pages known to be missing,
but the page numbers are unknown.-->
<!ELEMENT missing-pages  (#PCDATA) >

<!--The first (begin) and last (end) pages with bibliographic information.
Normally, begin will always = 1.-->
<!ELEMENT bib-pages  (begin,end) >

<!--The first and last pages on which the abstract appears. For US documents,
the abstract normally begins on page 1.-->
<!ELEMENT abstract-pages  (begin,end) >

<!--The first and last page numbers of the drawing pages. In US documents,
drawings normally follow the abstract and preceed the description. Drawing
pages do not overlap with the preceeding or following subdocuments.-->
<!ELEMENT drawings-pages  (begin,end) >

<!--The first and last pages of the description. The last page of the
description might be the same as the first page of the claims. Sequence
listings are normally between the description and the claims.-->
<!ELEMENT description-pages  (begin,end) >

<!--The first and last page of the claims. The first page of claims might be
the same as the last page of the description. Sequence listings are usually
between the description and the claims.-->
<!ELEMENT claims-pages  (begin,end) >

<!--The first page of the first certificate of correction and the last page of
the last certificate of correction.-->
<!ELEMENT certificate-of-correction-pages  (begin,end) >

<!--The first page of the first reexamination certificate and the last page of
the last reexamination certificate.-->
<!ELEMENT reexamination-pages  (begin,end) >

<!--First image page on which there is any part of the subdocument in
question.-->
<!ELEMENT begin  (#PCDATA) >

<!--Last image page on which there is any part of the document in question.-->
<!ELEMENT end  (#PCDATA) >

<!--Country: use ST.3 country code, e.g. DE, FR, GB, NL, etc. Also includes
EP, WO, and other regional authorities.
ST.32 name: B190; B330-->
<!ELEMENT country  (#PCDATA) >

<!--The number of the referenced patent (or application) document.
```

```
ST.32 name: B110; B210; B310-->
<!ELEMENT doc-number  (#PCDATA) >

<!--Document kind code; e.g. A1 Kind codes changed effective 2001-01-02 to
accommodate pre-grant publication status. A1 - Utility Patent Application
published on or after January 2, 2001. A2 - Second or subsequent publication
of a Utility Patent Application. A9 - Corrected published Utility Patent
Application. Bn - Reexamination Certificate issued prior to January 2, 2001.
NOTE: "n" represents a value 1 through 9. B1 - Utility Patent (no pre-grant
publication) issued on or after January 2, 2001. B2 - Utility Patent (with
pre-grant publication) issued on or after January 2, 2001. Cn - Reexamination
Certificate issued on or after January 2, 2001. NOTE: "n" represents a value 1
through 9 denoting the publication level. E1 - Reissue Patent. Fn -
Reexamination Certificate of a Reissue Patent NOTE: "n" represents a value 1
through 9 denoting the publication level. H1 - Statutory Invention
Registration (SIR) Patent Documents. SIR documents began with the December 3,
1985 issue. I1 - "X" Patents issued from July 31, 1790 to July 13, 1836. I2 -
"X" Reissue Patents issued from July 31, 1790 to July 13, 1836. I3 -
Additional Improvements - Patents issued between 1838 and 1861. I4 - Defensive
Publication - Documents issued from November 5, 1968 through May 5, 1987. I5 -
Trial Voluntary Protest Program (TVPP) Patent Documents. NP - Non-Patent
Literature. P1 - Plant Patent issued prior to January 2, 2001. P1 - Plant
Patent Application published on or after January 2, 2001. P2 - Plant Patent
(no pre-grant publication) issued on or after January 2, 2001. P3 - Plant
Patent (with pre-grant publication) issued on or after January 2, 2001. P4 -
Second or subsequent publication of a Plant Patent Application. P9 -
Correction publication of a Plant Patent Application. S1 - Design Patent.
ST.32 name: B130-->
<!ELEMENT kind  (#PCDATA) >

<!ELEMENT name  (#PCDATA) >

<!ATTLIST name
               name-type (legal | natural)   #IMPLIED  >

<!--Format: YYYYMMDD-->
```

<!ELEMENT date (#PCDATA) >

5 Examples

Patent Example:

A root directory listing for a DLT grants and rescans published the 12th week of 2002:

```
G00001.tid
20020319.contents
2002-12
```

A directory listing for new documents published in the 12th week of 2002, showing the subdirectory for document 6,342,021:

```
2002-12
   |-06
   |  |--245
   |  |  |--001
   |  |  |--002
   …  …  …
   |  |--342
   |  |  |--021
   |  |  |  |--00000001.tif
   |  |  |  |--00000002.tif
   …  …  …  …  …
   |  |  |  |--00003999.tif
   |  |  |  |--us-patent-image.xml
   |  |  |--022
   …  …  …  …
```

A contents list for documents published the week of 2002-03-19:

```
06245001
06245002
…
06342022
06342023
```

Application Example:

A tape which contains a range of publication dates, January 1, 2002 to February 15, 2002, generated on 18 September, 2002:

```
T00005.tid
20020918.contents
P20020101-20020215
```

A directory listing for applications with issue dates between January 1, 2002 to February 15, 2002, showing the subdirectory for document US20020005880A1:

```
P20020101-20020215
   |-US
   |  |--2002
   |  |  |--0000
   |  |  |--0001
   …  …  …
   |  |  |
```

```
|  |  |--0005
|  |  |  |--001
|  |  |  |--002
…  …  …  …  …
|  |  |  |--880
|  |  |  |  |--A1
|  |  |  |  |  |--00000001.tif
|  |  |  |  |  |--00000002.tif
…  …  …  …  …  …
|  |  |  |  |  |--00000023.tif
|  |  |  |  |  |--us-patent-image.xml
|  |  |  |--881
|  |  |  |  |--A1
…  …  …  …  …  …
```

A contents list for documents published between 2002-01-01 and 2002-02-15:

```
US20020000001A1
US20020000002A1
…
US20020019998A1
US20020019999A1
```